Supplementary Material

A. Failure cases in visual question answering

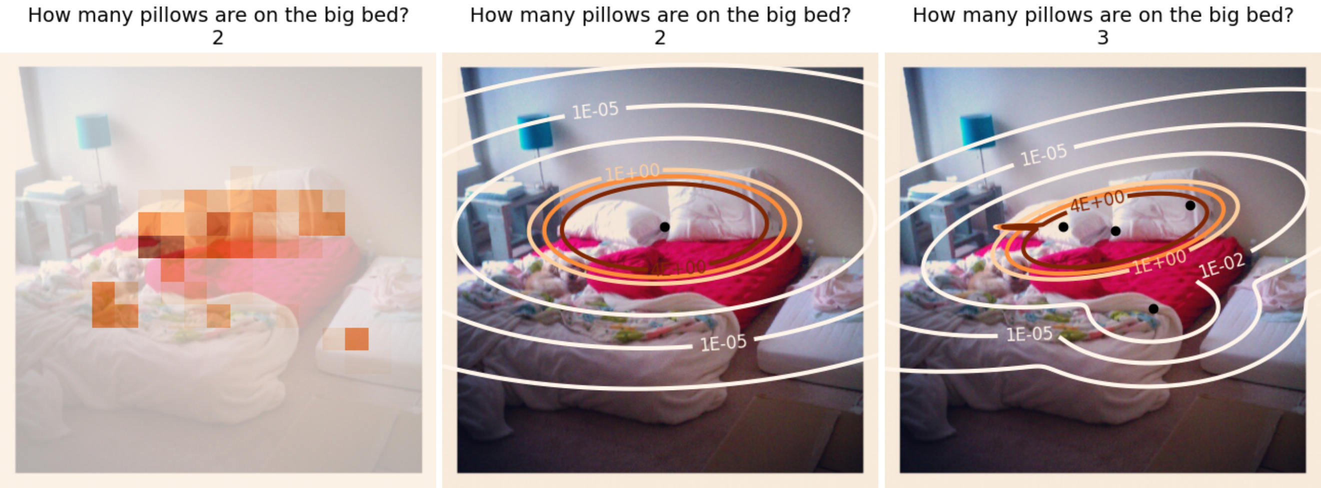
How many pillows are on the big bed?
2
How many pillows are on the big bed?
2
1E-05
1E+00
1E+00
1E-05
How many pillows are on the big bed?
3
1E-05
4E+00
1E+00
1E-02
1E-05

Figure 8. Examples of attention maps in the VQA-v2 dataset. Left: discrete softmax attention. Middle: unimodal continuous attention. Right: Multimodal continuous attention (ours).

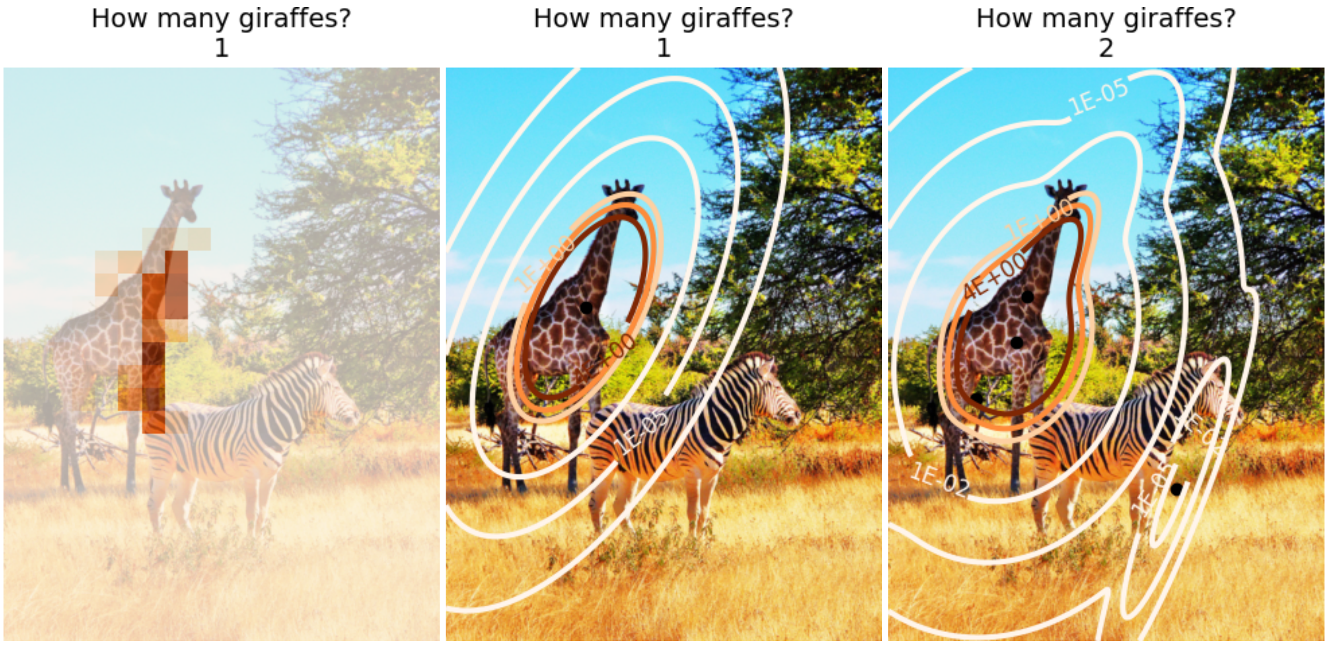

Figure 9. Examples of attention maps in the VQA-v2 dataset. Left: discrete softmax attention. Middle: unimodal continuous attention. Right: Multimodal continuous attention (ours).

In §6 we presented several attention maps generated by different models and discussed the main strengths of multimodal continuous attention, when compared to discrete or unimodal continuous attention. Although our model tends to perform considerably better in complex situations where it is possible to identify multiple regions of interest or a single region with a complex shape, there are cases in which fitting a multimodal distribution as the attention density may lead to incorrect answers. For instance, in the example in Figure 8, when looking for pillows on the bed, our model focuses on more than one region and possibly confuses the messy bed cover with a pillow. A similar situation is illustrated by the example in Figure 9, where the zebra is taken as being another giraffe. These examples suggest that in spite of being able to generate unimodal attention maps when the relevant regions in the image are contiguous or unique, our model sometimes fails as a result of its capability of looking for more than one region in the image.